# 面向形态丰富语言的翻译规则选择方法

王志洋 吕雅娟 孙萌 姜文斌 刘群

**摘要：**目前的机器翻译模型都是针对形态变化简单的语言（如英语）设计的，不太适合于形态丰富语言（如维吾尔语）。在本文中，我们通过区别对待形态丰富语言中的词干与词缀，提出了一种新型的面向形态丰富语言的翻译规则选择方法。我们用词干作为基本翻译单元以缓解数据稀疏问题，此外，每条词干粒度的翻译规则上还附着一个词缀分布。在翻译时，通过计算待翻译片段的词缀分布与翻译规则词缀分布的相似度，来选择更合适的翻译规则。从三种形态丰富语言（维吾尔语、哈萨克语、柯尔克孜语）到汉语的翻译实验表明，该方法显著改善了翻译质量。

**关键词：**机器翻译 形态丰富语言 词缀分布 相似度 动态特征



## 1 引言

形态丰富语言是指词的形态变化比较复杂、丰富的一类语言。从形态学角度来说，语言可以分为孤立语、屈折语、黏着语和多式综合语。实际上，除了孤立语和少数屈折语，绝大部分语言都属于形态丰富语言。我国少数民族语言中的维吾尔语、蒙古语等，以及我国周边绝大部分国家的官方语言都属于此类。

表1. 形态变化类型示例

| 变化类型 | 示　　例 |
|---|---|
| 屈折变化 | gül (花) : gül+üm (我的花) |
| | kitab (书) : kitab+ing (你的书) |
| | doppa (帽子) : doppa+si (他的帽子) |
| 一致性 | **män** gezit oqu+y+**män** (我读报纸) |
| | **sän** gezit oqu+y+**sän** (你读报纸) |
| 复合变化 | qar (雪), leyle (花) : qarleylesi (雪莲花) |
| | tax(石头), paqa (青蛙) : taxpaqa (乌龟) |
| 语音和谐 | markt**ep** (学校) : markt**i**pim (我的学校) |
| | chir**aq** (台灯) : chirqing (你的台灯) |

形态丰富语言最显著的特点是形态变化复杂。下面以维吾尔语[1]为例，来说明这类语言的形态变化特点。表 1 列举了形态丰富语言中常见的形态变化方式。**屈折变化**指的是通过在词干上加接词缀，导致其语法功能改变，同时也改变了单词的拼写。如在名词 doppa（帽子）后缀接第三人称单数后缀“si”，就变成了 doppasi（他的帽子）。**一致性**指的是句子或短语的不同部分存在对应关系。为了与相应的语法关系一致，需要改变词形以保持一致性。当表达“我读报纸”时，需要在动词 oqu（读）的后面加上表示第一人称单数的词缀“män”，以保持一致性。此外，还有**复合变化**。两个词连接在一起可以生成表达不同意思的新词。如名词 tax（石头）和 paqa（青蛙）复合构成词 taxpaqa，表示的意思是“乌龟”。**语音和谐**是表音文字中常见的现象。不同的音节组合在一起时，部分字母需要发生一定的变化（增音、脱落、弱化等），以符合发音规律。这一系列丰富的形态变化方式，可能会导致一个词干生成成百上千种新的词形。如果将每一个词形都单独建模成词，会导致严重的数据稀疏问题。这对传统的统计机器翻译模型是一个巨大的挑战。

虽然各国很早就展开了机器翻译方面的研究，但主要是围绕英语等语言进行，针对形态丰富语言的研究较少。在涉及形态丰富语言到汉语翻译的研究中，更多的是沿用之前在英语

[1]为了书写和阅读上的方便，本文一律使用拉丁字母来表示形态丰富语言。

等语言的翻译上表现良好的方法。但由于形态丰富语言自身的特点，翻译效果并不尽如人意。此外，目前最为成功的统计机器翻译方法需要大规模的双语平行语料库作为训练语料，而对于形态丰富语言和汉语间的翻译来说，由于缺乏大规模的双语平行语料资源，单纯的统计方法可能很难取得理想的效果。另一方面，大部分形态丰富语言的已有的语言处理基础相对薄弱，研究工作较少，缺乏实用的词法分析（也称为形态分析）、句法分析等工具，用于词法分析和句法分析的标注语料库也十分有限。

我国是一个拥有 56 个民族的多元文化共存的国家。除汉族以外，少数民族中的维吾尔、蒙古、哈萨克等民族也都有自己的文字，并在本民族广泛使用。其中维吾尔语、蒙古语、哈萨克语等都属于形态丰富语言。在中国周边的 21 个国家中，大部分国家的官方语言形态变化都比较丰富，如俄语、日语、朝鲜语、印尼语、马来语、印地语（其中俄语、朝鲜语同时也是我国的少数民族语言）等。在我国的 21 个邻国中，有 16 个国家全部使用或部分使用形态丰富语言作为官方语言，比例高达 76%。因此，研究形态丰富语言到汉语的翻译有其现实意义。通过研究形态丰富语言到汉语之间的机器翻译，可以促进区域间的多元文化交流，加强经济、文化、教育等多个领域的合作。

在接下来的章节中，我们首先描述形态丰富语言翻译的国内外研究现状（§2）；然后具体介绍基于词缀消歧的翻译规则选择的方法（§3）。在模型介绍完毕之后，§4 详细描述和分析了实验结果，最后给出总结和展望（§5）。



## 2 国内外研究现状

在大多数自然语言处理任务中，词都作为知识表示的原子单元。在统计机器翻译中，也将词看作是原子翻译单元，而不考虑词内部的形态构成。从起始的基于词的翻译模型[1]，到之后改进的短语模型[2]、层次短语模型[3]以及句法模型[4-6]，都保留了这种假定。在存在较大双语语料库的前提下，这些改进模型在翻译孤立语（如汉语）和形态变化不太丰富的语言（如英语）时，效果很不错。但对形态变化丰富的语言来说，一个词干可以缀接多个词缀（前缀或者后缀），这将会生成成百上千种新的词形（surface form）。如果将词干相同的每个词形都单独建模成词，数据稀疏现象将会非常严重。如蒙古语动词词根“UILED”（做），理论上至少有一千七百多种变化形式[7]。

形态丰富语言翻译有三种不同的翻译粒度。一种是词（word），即使词干相同的词形，也单独建模成词。使用词粒度翻译，可以抽取更准确的翻译规则。但在语料规模不大的前提下，数据稀疏问题将严重影响对齐和翻译质量。另外一种是词干（stem），词干是词除去构形词缀的部分，表达了词的基本意义。词干粒度的翻译规则覆盖率更大，但毕竟丢弃了一些有用的词缀，规则会存在歧义问题。最后一种粒度是词素（morpheme），词素是构词的最小有意义单位。将构成词的每个词素都作为单独的翻译单元，构成句子的元素个数将会增加，给词语对齐和翻译解码带来负担。

学术界很早就展开了形态丰富语言翻译（涉及的语言有德语、西班牙语、阿拉伯语、印地语、捷克语、芬兰语等）的相关研究。形态丰富语言翻译的相关研究可以分为三类。

第一类是针对数据稀疏的问题，通过形态分析，对形态丰富语言进行预处理，以提高翻译质量。戈德华特（Goldwater）和麦克洛斯基（McClosky）[8]尝试多种词素组合策略来表示捷克语，改善了捷克语到英语的翻译质量；波波维奇（Popovic）和奈伊（Ney）[9]通过将西班牙语种的形容词用其词根替换，将所有塞维利亚语的单词用词根替换，在语料受限的情况下，使西班牙语到英语、塞尔维亚语到英语的翻译质量有了明显的提升。哈巴什（Habash）

和萨达特（Sadat）[10]在阿拉伯语到英语的翻译中，在预处理中使用了不同的形态分离策略。实验表明，形态分离并不是分得越细越好，需要根据实验来确定一个合适的翻译粒度。李（Lee）[11]引入双语信息，选择合适的粒度来表示输入，平衡两种语言间的词形变化差异问题； 杨（Yang）和基尔霍夫（Kirchhoff）[12]的工作中，当遇到未登录词（Out Of Vocabulary，简称 OOV）时，将其退化到词干进行翻译；有一些研究还针对复合变化，通过分解复合词[13]来改善翻译；还有一部分相关工作就是扩展输入信息，如使用词图结构（lattice）[14]、复述（paraphrase）[15]等进行容错翻译。

另外一类是充分利用形态和句法信息，联合多种要素来指导翻译。代表性的是开源翻译系统 Moses 中的基于要素（factor）的模型[16]。该模型在生成目标译文的同时生成相应的词性，词形变化等信息，然后利用高阶的词性 N-gram 模型[2]以及词形的因子化 N-gram 模型优化目标词的选择。词性（POS）标注、格（case）、甚至超级标注（super tag）[17]，都可以当作要素加入到模型中，以提高翻译效果。但对绝大部分形态丰富语言而言，高质量的处理工具（如词性标注工具、CCG 句法分析工具等）目前都还无法获得。还有一部分工作更进一步通过句法分析，对源语言（形态丰富语言）进行预调序，让其词序更符合目标语言，代表性工作有引文[18][19]等。此外，为了克服语言间的词形变化差异现象，耶尼特里茨（Yeniterzi）和奥夫拉泽尔（Oflazer）[20]尝试通过对英语的句法分析，重组英语端，让其更类似土耳其语，以完成英语到土耳其语的翻译。拉曼纳森（Ramanathan）等人[21]深入挖掘英语端的相应知识，并将其映射至印地语端,来改善英语到印地语的翻译。这类方法可以很大地改善翻译效果，但前提是必须要有相应的句法分析工具可供使用。

第三类研究则力图克服多数形态丰富语言双语平行语料资源匮乏带来的困难。国际上常见的做法是利用相似语言的资源，或者使用桥接语言（pivot language）来进行翻译[22][23]。但对大部分形态丰富语言而言，这类资源也同样非常缺乏。因此，借鉴相似语言资源和采用桥接语言的做法也不太适用。

总的来说，目前形态丰富语言翻译的研究工作，主要面向的是资源不太匮乏的语言，借助一些语言处理工具，像词法和句法分析工具等来改善翻译效果。但实际上，绝大部分形态丰富语言的双语资源都有限，而且缺乏相应语言处理工具。

# 3 基于词缀消歧的翻译规则选择方法

词缀，尤其是构形词缀（inflectional affix），表达了所属词的语法意义，如人称、时态、数的变化以及格变化等，这些对于准确地描述翻译规则具有重要的作用。因此，我们在抽取词干粒度的翻译规则的时候，同时保留相应的词缀信息。

## 3.1 翻译规则表示

图 1（B）中给出了两条维吾尔语到汉语翻译规则示例。翻译规则实例（1）（3）相同，表示从不同的双语句对中抽取出相同的规则实例。其中词缀分布使用经典的向量空间模型（Vector Space Model，简称 VSM）来表示。从图中我们可以看到，尽管这两条翻译规则的源端是一样的，但它们的词缀分布情况却差别很大。第一条翻译规则中的后缀“gha”是维语中的向格，表示所属关系，类似于英语中的介词“of”；第二条翻译规则中的后缀“da”是时位格，表示的是位置信息。这两个词缀的区别也直接反映在目标短语上。因此，当待翻

[2] 大词汇连续语音识别中常用的一种语言模型

译的片断为“zunyi/STM yihin/STM+i/SUF+da/SUF/+...”[3]时，在源端词干序列都匹配上的前提下，我们倾向于模型选择第二条翻译规则。我们可以通过计算待翻译片断和候选规则的词缀分布的相似度来鼓励选择更合适的目标规则。

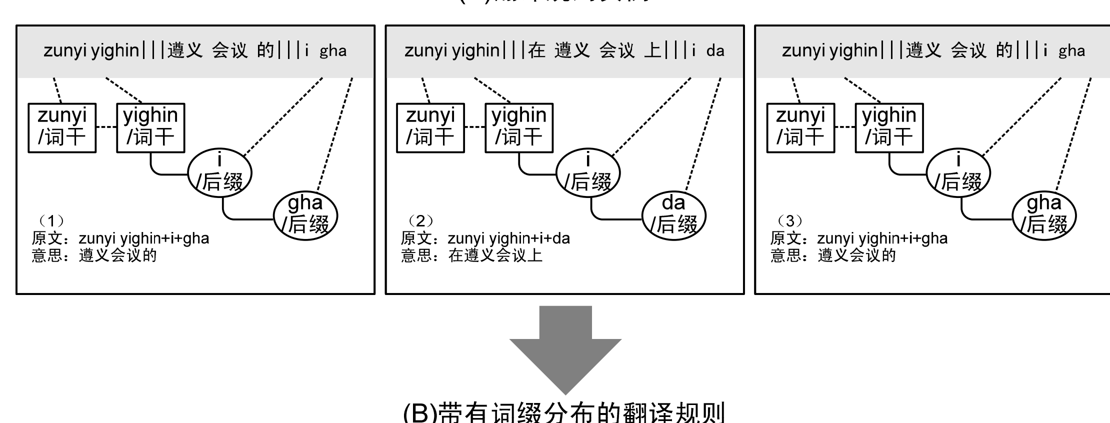



图1.　词缀分布抽取和翻译规则表示

## 3.2 规则抽取与参数估计

翻译规则抽取的流程如下：

1. 源语言端表示为词干（维吾尔语），目标语言端仍为单词（汉语）。然后进行对齐和规则抽取，最终获得的是词干-词粒度的翻译规则和对应的概率得分。

2. 源语言端表示为词干+词缀组合的形式，目标语言端为单词。使用步骤 1 中词干-词粒度的对齐结果（前文中提到维吾尔语中每个单词只包含一个词干），进行词干粒度的规则抽取，同时将相应的词缀信息保留在规则实例中。

3. 在步骤 2 抽取的规则实例的基础上，利用向量空间模型对词缀分布的参数进行估计（详细见下文），以获得每条规则的词缀分布情况。

4. 将步骤 1 和步骤 3 的翻译规则进行合并，主要是将词缀分布加入到原始的翻译规则表中，从而得到最终的翻译规则。

如前所述，词缀分布表示为向量的形式。这里我们重点阐述一下如何得到词缀分布的向量表示。我们将具有相同源端的翻译规则看作是一个“文档集合”，这样“集合”内的每一条翻译规则就是一个“文档”。我们的目标就是利用词缀分布信息将每个“文档”分类到对应的目标短语。具体可以分为以下三步：

首先，在抽取词干粒度翻译规则的同时，保留相应的词缀信息。在图 1（A）中，从维吾尔语的原始形式可以看到，相应的词干序列构成翻译规则的源端，剩下的词缀序列及其计数也保留下来。

然后，源端相同的规则构成一个集合，在这个集合内，我们可以使用经典的 TF-IDF[4]来

---

[3] 其中，STM 表示词干，SUF 表示后缀

[4] term frequency–inverse document frequency,一种用于资讯检索与资讯探勘的常用加权技术

表示相关词缀的权重。

最后，在同一个集合内，我们需要将目标端也进行相同的翻译规则聚合，这里我们采用基于质心的分类算法[24]来表示最终的词缀分布结果：

$$\boldsymbol{d}_{\text{rule}} = \frac{1}{\boldsymbol{N}} \sum_{i \in N} \boldsymbol{d}_i$$

其中，$\boldsymbol{N}$ 表示具有相同目标端的规则数目，$\boldsymbol{d}_{\text{rule}}$ 是通过平均目标端相同的词缀分布得到。这样，对于待翻译的片断，我们首先通过形态分析获得其词干序列和词缀分布（表示为向量）。其中，词干序列用来检索翻译规则表以获得翻译候选。当源端词干序列匹配成功后，我们再计算待翻译片断和候选翻译规则的词缀分布的相似度。在本文中，相似度 **sim** 通过向量的夹角余弦来衡量：

$$\boldsymbol{sim}(\boldsymbol{d}, \boldsymbol{d}_{\text{rule}}) = \frac{\boldsymbol{d} \cdot \boldsymbol{d}_{\text{rule}}}{|\boldsymbol{d}| \times |\boldsymbol{d}_{\text{rule}}|}$$

词缀分布的相似度得分将作为一个动态特征加入到对数线性模型（log-linear model）[25]中，以指导解码器选择更合适的翻译规则。

## 3.3 选取有效的词缀

词缀分布在解码器选择更合适的翻译规则时作用显著。但是，词缀往往都是通过对单语的形态分析获得的，这样得到的词缀集合不一定适合于机器翻译。直觉上来说，如果同时考虑目标语言端信息，使用双语来约束词缀的生成，可能会得到更适合机器翻译的词缀集合。为了获得更有用的词缀，丢弃无用的词缀（类似于文本分类中的停用词表 stop list），我们提出了一种获得合适词缀集合的判别式方法。

给定词素粒度的对齐结果，我们可以确定对当前词缀如何操作。如果当前词缀和之前的词缀对齐到同一个目标词，这两个词缀应该合并成一个词缀（称之为 merge）；如果当前词缀和之前的词缀对齐到不同的目标词，则应该单独保留（称之为 keep）；当其对齐到空时，则应该删除（称之为 delete）。

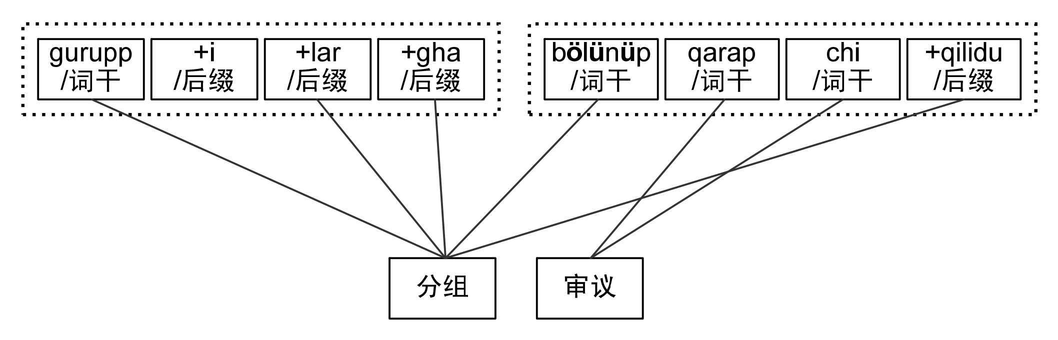



图2. 维吾尔语-汉语对齐示例

在图 2 中，后缀“gha”和它之前的后缀“lar”对齐到同一个目标词，两者应该合并构成一个新的词缀；后缀“qilidu”和之前的词素对齐到不同的目标词，应该保留；而“i”对齐到空，应该直接删除。也就是说，在同一个词内，其组成词缀都可以分为三类：合并、保留、删除。分类实例可以直接从词素粒度的对齐语料上获得。为了获得分类模型，我们选择条件随机场（Conditional Random Field，CRF）[26]来对实例进行训练。CRF 是一种判别式概率模型，给定观察序列，可以计算得到输出状态序列的条件概率，常用于解决序列标注问题。该

模型不需隐马尔可夫模型(HMM)苛刻的独立性假设[27]，可以融合任意的特征。而且，不存在最大熵模型[28]的标记偏见问题，其求解的是当前观察序列的全局最优输出状态的条件概率。

表2. 词缀选择的特征模板和实例

| 特征模板 | 实　　例 |
|---|---|
| $C_i$ ($i$=-2,⋯,2) | $C_{-2}$=+I, $C_{-1}$=+lar, $C_0$=+gha, $C_1$=bölünüp, $C_2$=qarap |
| $P_i$ ($i$=-2,⋯,2) | $P_{-2}$=M, $P_{-1}$=M, $P_0$=E, $P_1$=S, $P_2$=S |
| $C_iC_{i+1}$($i$=-2,⋯,1) | $C_{-2}C_{-1}$=+i+lar, $C_{-1}C_0$=+lar+gha, $C_0C_1$=+ghabölünüp, $C_1C_2$= bölünüp qarap |
| $P_iP_{i+1}$($i$=-2,⋯,1) | $P_{-2}P_{-1}$=MM, $P_{-1}P_0$=ME, $P_0P_1$=ES, $P_1P_2$=SS |

具体而言，本实验中使用的 CRF 工具是开源的软件 CRF++[5]。表 2 是训练分类模型使用的特征模板。除了邻居窗口的词素特征外，我们还使用了词素在词中的位置信息（词首、词中、词尾和单独成词）。引入位置信息的主要目的是为了保留词的内部结构信息。假定当前考虑的是图 2 中的后缀“gha”。用 B，M，E，S 分别表示词素在词中的位置信息：词的开始，词的中间，词的结尾，以及单独成词。为了获得更好的分类效果，本方法可以迭代训练。

## 4　实验

为了验证提出方法的效果，我们在三组语言对上进行了翻译实验：维吾尔语-汉语，哈萨克语-汉语，柯尔克孜语-汉语。其中，维吾尔语、哈萨克语和柯尔克孜语都是在我国西部地区使用较多的少数民族语言，均属于阿尔泰语系突厥语族，形态变化都异常丰富。相关语料来源于全国机器翻译研讨会（China Workshop of Machine Translation，简称 CWMT[6]）的翻译评测。需要说明的是，由于 CWMT 的相关评测属于进展测试（progress tests），我们得不到评测中所用测试集合，这里的测试集合是我们自己构造的。表 3 是语料的统计信息，其中“*”号后的数字表示参考译文的个数。从表中可以看到，经过形态分析之后，三种形态丰富语言的词汇量都减少很多，缓解了数据稀疏问题。

表3. 数据集合的统计信息

| 数据集合 | 句对数 | 词汇数 | | | 单词数 | | |
|---|---|---|---|---|---|---|---|
| | | 词 | 词干 | 词素 | 词 | 词干 | 词素 |
| 维汉训练集 | 50K | 69K | 39K | 42K | 1.2M | 1.2M | 1.6M |
| 维汉开发集 | 0.7K*4 | 5.9K | 4.1K | 4.6K | 18K | 18K | 23.5K |
| 维汉测试集 | 0.7K*1 | 4.7K | 3.3K | 3.8K | 14K | 14K | 17.8K |
| 哈汉训练集 | 50K | 62K | 40K | 42K | 1.1M | 1.1M | 1.3M |
| 哈汉开发集 | 0.7K*4 | 5.3K | 4.2K | 4.5K | 15K | 15K | 18K |
| 哈汉测试集 | 0.2K*1 | 2.6K | 2.0K | 2.3K | 8.6K | 8.6K | 10.8K |
| 柯汉训练集 | 50K | 53K | 27K | 31K | 1.2M | 1.2M | 1.5M |
| 柯汉开发集 | 0.5K*4 | 4.1K | 3.1K | 3.5K | 12K | 12K | 15K |
| 柯汉测试集 | 0.2K*4 | 2.2K | 1.8K | 2.1K | 4.7K | 4.7K | 5.8K |

对于语言模型，我们使用 SRI[7]的语言模型训练工具 SRILM[29]，根据训练语料的目标端

[5] http://crfpp.sourceforge.net/
[6] http://mt.xmu.edu.cn/cwmt2011/en/index.html
[7] http://www.speech.sri.com/projects/srilm/

训练 5 元语言模型；Moses[8]的短语系统作为基线系统，系统的特征权重使用最小错误率算法[30]来调参，目标是使词级 BLEU[9]值最大化[31]。为了能够动态地将相似度特征加入到对数线性模型中，我们在 Moses 短语系统的基础上重新构筑了可动态计算词缀分布相似度的解码器。

之前也提到，对于绝大部分形态丰富语言，语料和工具资源都相对匮乏，相应的高质量的形态分析工具很难获得；因此，这里我们使用无监督形态分析方法对所用语言进行形态分析，以更好地验证本方法与具体语言的无关的特性。跟文献[32]类似，我们也采用芬兰赫尔辛基大学开发的无监督分析工具 Morfessor[10]，这里为了模拟资源匮乏语言，我们没有对语音和谐现象进行还原处理。Morfessor 是根据最小描述长度（Minimum Description Length，简称 MDL）来生成形态分类析结果，文献[33]将 Morfessor 生成的“词素”（文中称为 morph，无监督最小切分单位；和语言学意义上的词素有差别）分为三类：前缀（prefix，PRE）、词干（stem，STM）和后缀（suffix，SUF），据此我们来区分词干和词缀。在实验中，我们选择训练语料库中出现次数最多的前 5000 词来训练 Morfessor 的切分模型。

## 4.1 实验结果与分析

表 4 是三种突厥语族语言到汉语的翻译结果。其中词（*word*）方法是基线系统，将词粒度作为原子翻译单位；词干（*stem*）方法表示在翻译时，词用对应的词干代替；词素（*morph*）方法表示使用词素作为最小翻译单位。词缀（*affix*）方法对应本文提出的用词干翻译，用词缀分布进行规则选择的方法；CRF-词缀方法（*crf-affix*）是在词缀的据 CRF 模型选择了更有用词缀后的结果。黑体部分表示和基线系统相比，实验结果具有显著性[34]提高。从表中可以看出，对三种语言，使用词干作为最小翻译单位的效果均好于使用词和词素；而我们提出的方法表现也均好于词干的翻译效果。

表4. 维吾尔语、哈萨克语、柯尔克孜语到汉语的翻译结果

| | UY-CH(%) | KA-CH(%) | KI-CH(%) |
|---|---|---|---|
| 词(word) | $31.74_{+0.0}$ | $28.64_{+0.0}$ | $35.05_{+0.0}$ |
| 词干(stem) | $\mathbf{33.74}_{+2.0}$ | $\mathbf{30.14}_{+1.5}$ | $35.52_{+0.47}$ |
| 词素(morph) | $32.69_{+0.95}$ | $29.21_{+0.57}$ | $34.97_{-0.09}$ |
| 词缀(affix) | $\mathbf{34.34}_{+2.6}$ | $\mathbf{30.19}_{+2.27}$ | $35.96_{+0.91}$ |
| CRF 词缀法 | $\mathbf{34.64}_{+2.9}$ | $\mathbf{31.24}_{+2.60}$ | $\mathbf{36.27}_{+1.22}$ |

具体来说，在维吾尔语到汉语的翻译任务中，和基线系统相比，CRF 词缀方法 BLEU 值提高了 2.9 个百分点；而且与词干粒度翻译相比，也有 0.9 个百分点的提高。哈萨克语到汉语的翻译中，提升效果也比较明显，相对基线系统和词干翻译，分别有 2.6 和 1.1 个百分点的 BLEU 值上的提高。此外，使用 CRF 模型选择更有用的词缀过后，和不处理之前也有 0.33 个百分点的提高。柯尔克孜语翻译到汉语时，较前两组语言对相比，提升幅度稍小，但也有 1.22 个百分点的提高。使用 CRF 模型来选择词缀集合过后，在三种语言上都带来了一定的翻译质量的提高；但总的来说，提高幅度不是很大。一个可能的原因是，词缀是合并、保留还是删除，依赖于词素粒度的对齐，尤其是词缀本身的对齐结果，而该对齐效果往往不尽如人意。作为下一步工作，我们希望先通过改善词缀的对齐效果，得到更准确的词缀分类实例，以改善词缀集合选择的结果。

通过观察分析翻译结果，我们发现，和基线系统相比，我们的模型生成的翻译结果更流利。具体地，改善效果主要体现在两个方面：

8 http://www.statmt.org/moses/
9 Bilingual Evaluation Understudy，由 IBM 于 2002 年提出的一种机器翻译质量自动评测方法
10 http://www.cis.hut.fi/projects/morpho/

- **降低了未登录词（OOV）的比例**由于我们使用词干作为原子翻译单位，具有相同词干的词形都使用其词干来表示，从而很好地缓解了数据稀疏问题。在表 5 例 1 中，虽然我们的训练语料库中没有词“qutquzishi”，但存在很多以“qutquz”为词干的词形。因此，在使用词干粒度翻译时，“qutquzishi”就变成“qutquz”，从而能翻译出来。可以看到，词干粒度翻译可以明显地降低未登录词的比例。

- **选择更合适的词汇** 在表 5 的两个例子中，引入词缀分布来消歧，可以选择更合适的词汇，生成的翻译结果更符合语法。例 1 中生成了相匹配的量词“名”，例 2 中则包含了对应的介词“向”。

表5. 维吾尔语到汉语的翻译结果示例

| | |
|---|---|
| 例 1 原文 | munasiwetlik tarmaqlarning pütün küchi bilen qutquzishi arqiliq , 1400 din artuq yoluchi qutquzwëlindi . |
| 参考译文 | 经过 全力 救援 成功 解救 出 一千四百 多 名 被 困 人员 。 |
| 词（word） | 有关 部门 全力 qutquzishi ， 旅客 qutquzwëlindi 。 |
| 词干（stem） | 有关 部门 全力 **营救** ， 1400 多 旅客 **救** 出 。 |
| 后缀（affix） | 有关 部门 全力 **营救** ， 1400 多 **名** 乘客 **救** 出 。 |
| 例 2 原文 | hemde qurbanlarning tughqanliridin hal soridi . |
| 参考译文 | 并 向 烈士 亲属 表示 深切 慰问 。 |
| 词（word） | 并 烈士 失去 亲人 的 慰问 。 |
| 词干（stem） | 并 烈士 亲属 表示 慰问 。 |
| 后缀（affix） | 并 **向** 烈士 亲属 表示 慰问 。 |

## 4.2 词法分析质量的影响

以上实验均在无监督形态分析结果上进行，都有效改善了翻译效果。更进一步，我们想验证一下形态分析的质量会对该方法产生何种影响。我们使用文献[35]中提出的方法构建了一个基于标注语料的维吾尔语词法分析工具，并测试其翻译效果。图 3 是和无监督分析工具 Morfessor 的结果对比。可以看到，利用有监督的词法分析工具获得词干、词缀后，除了词素粒度的结果，使用词干和增加词缀消歧的模块的结果均好于无监督分析后的翻译结果，约有 0.2 到 0.38 个百分点的提高。这说明，我们提出的方法在形态分析质量改善时，翻译效果也随之改善。表 6 给出了两种形态分析方法的结果在统计上的差别。有监督的分析方法生成的词素类型更少，尤其是词缀，只有 Morfessor 的 1/10。生成的词缀往往更具有语法意义，能更好地指导翻译规则的选择。

表6. 无监督和有监督词法分析后语料统计信息对比

| UY | | 无监督切分 | 有监督切分 |
|---|---|---|---|
| 词干 | 词汇数 | 39K | 21K |
| | 单词数 | 1.2M | 1.2M |
| 后缀 | 词汇数 | 3.0K | 0.3K |
| | 单词数 | 0.4M | 0.7M |

## 4.3 语料规模的影响

此外，我们还在一个较大规模的维吾尔语-汉语平行语料库上进行了实验，以验证我们的方法在相对大规模语料库上的有效性。我们将收集的约 30 万关于政府新闻的维汉语料库随机划分为六个部分：5 万、10 万、15 万、20 万、25 万和 30 万，来验证语料规模的大小对翻译效果的影响。仍然沿用之前的开发集和测试集，并确保与现有的训练集没有重叠。图

4 是翻译曲线。可以看到，不论语料库规模如何，基于词缀消歧的词干翻译方法都始终表现最好：在语料库规模较小时，该方法提高很明显；随着语料库规模的加大，提升幅度稍微减小。即使如此，在使用全部的 30 万的双语语料时，该方法仍然有 0.7 个百分点的提高。

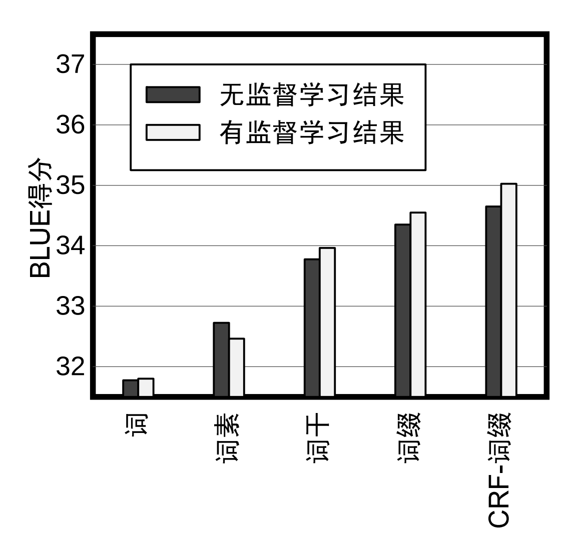



图3. 无监督和有监督词法分析对翻译结果的影响

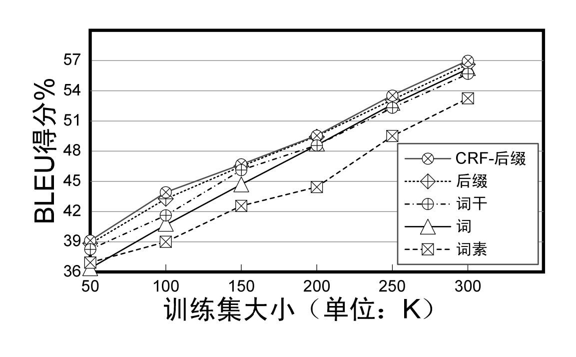



图4. 不同语料规模对翻译结果的影响

## 5 总结与展望

本文通过区别对待词干与词缀，提出了一种新颖的面向形态丰富语言的翻译规则选择方法。在整个翻译流程中，我们使用词干作为原子翻译单元。此外，每条词干粒度的规则都会附带一组相应的词缀分布，通过计算其与待翻译片段的词缀分布的相似度，来帮助解码系统选择更合适的翻译规则。在三种不同形态丰富语言上的实验表明，该方法显著改善了翻译质量，尤其是在双语语料相对匮乏时，效果提升很明显。

本文研究了区别对待词干和词缀来对形态丰富语言进行翻译的方法，这样的工作尚属首次。该方法与具体的语言对无关。我们计划下一步在更多的形态丰富语言上来验证本文的结果，并改善翻译质量。此外，这里的词干类似于内容词，而词缀则对应功能词。按这种类比，我们的方法也应该可以用在翻译形态变化不太丰富的类似英语的语言上。

**参考文献：**

[1] Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: parameter estimation. Comput. Linguist., 19(2):263–311, 1993.

[2] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In Proceedings of NAACL, pages 48–54, 2003.

[3] David Chiang. A hierarchical phrase-based model for statistical machine translation. In Proceedings of ACL, pages 263–270, 2005.

[4] Chris Quirk, Arul Menezes, and Colin Cherry. Dependency treelet translation: syntactically informed phrasal SMT. In Proceedings of ACL, pages 271–279, 2005.

[5] Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and

Ignacio Thayer. Scalable inference and training of context-rich syntactic translation models. In Proceedings of COLING/ACL, pages 961–968, 2006.

[6] Yang Liu, Qun Liu, and Shouxun Lin. 2006. Treeto-string alignment template for statistical machine translation. In Proceedings of COLING-ACL, pages 609–616.

[7] 那顺乌日图，刘群，巴达玛放德斯尔. 面向机器翻译的蒙古语生成. 全国第六届计算语言学联合学术会议论文集, 2001.

[8] Sharon Goldwater and David McClosky. Improving statistical MT through morphological analysis. In Proceedings of HLT-EMNLP, pages 676–683, 2005.

[9] Maja Popovic and Hermann Ney. Statistical machine translation with a small amount of bilingual training data. In LREC workshop on Minority Language, pages 25–29, 2006.

[10] Nizar Habash and Fatiha Sadat. Arabic preprocessing schemes for statistical machine translation. In Proceedings of NAACL, Short Papers, pages 49–52, 2006.

[11] Young-Suk Lee. Morphological analysis for statistical machine translation. In Proceedings of HLT-NAACL, Short Papers, pages 57–60, 2004.

[12] Mei Yang and Katrin Kirchhoff. Phrase-based backoff models for machine translation of highly inflected languages. In Proceedings of EACL, pages 1017–1020, 2006.

[13] Philipp Koehn and Kevin Knight. Empirical methods for compound splitting. Proceedings of EACL, pages 187–193, 2003.

[14] Christopher Dyer, Smaranda Muresan, and Philip Resnik. Generalizing word lattice translation. In Proceedings of ACL: HLT, pages 1012–1020, 2008.

[15] Preslav Nakov and Hwee Tou Ng. Translating from morphologically complex languages: A paraphrase-based approach. In Proceedings of ACL: HLT, pages 1298–1307, 2011.

[16] Philipp Koehn and Hieu Hoang. Factored translation models. In Proceedings of EMNLPCoNLL, pages 868–876, 2007.

[17] Alexandra Birch, Miles Osborne, and Philipp Koehn. CCG supertags in factored statistical machine translation. In Proceedings of StatMT, pages 9–16, 2007.

[18] Peng Xu, Jaeho Kang, Michael Ringgaard, and Franz Josef Och. Using a dependency parser to improve smt for subject-object-verb languages. In Proceedings of NAACL, pages 245–253, 2009.

[19] Dmitriy Genzel. Automatically learning source-side reordering rules for large scale machine translation. In Proceedings of COLING, pages 376–384, 2010.

[20] Reyyan Yeniterzi and Kemal Oflazer. Syntax-to-morphology mapping in factored phrase-based statistical machine translation from English to Turkish. In Proceedings of ACL, pages 454–464, 2010.

[21] Ananthakrishnan Ramanathan, Hansraj Choudhary, Avishek Ghosh, and Pushpak Bhattacharyya. Case markers and morphology: Addressing the crux of the fluency problem in English-Hindi SMT. In Proceedings of ACL, pages 800–808, 2009.

[22] Preslav Nakov and Hwee Tou Ng. Improved statistical machine translation for resource-poor languages using related resource-rich languages. In Proceedings of EMNLP, pages 1358–1367, 2009.

[23] [23] Pidong Wang, Preslav Nakov, and Hwee Tou Ng. Source language adaptation for resource-poor machine translation. In Proceedings of EMNLP, pages 286–296, 2012.

[24] Eui-Hong Sam Han and George Karypis. Centroid-based document classification: Analysis experimental results. In Proceedings of PKDD, pages 424–431, 2000.

[25] Franz Josef Och and Hermann Ney. Improved statistical alignment models. In Proceedings of ACL, pages 440–447, 2000.

[26] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields:

Probabilistic models for segmenting and labeling sequence data. In Proceedings of ICML, pages 282–289, 2001.

[27] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In Proceedings of IEEE, pages 257–286, 1989.

[28] Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. A maximum entropy approach to natural language processing. Comput. Linguist., 22(1):39–71, 1996.

[29] Andreas Stolcke. SRILM - an extensible language modeling toolkit. In Proceedings of ICSLP, pages 311–318, 2002.

[30] Franz Josef Och. Minimum error rate training in statistical machine translation. In Proceedings of ACL, pages 160–167, 2003.

[31] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of ACL, pages 311–318, 2002.

[32] Sami Virpioja, Jaakko J. Väyrynen, Mathias Creutz, and Markus Sadeniemi. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In Proceedings of MT SUMMIT, pages 491–498, 2007.

[33] Mathias Creutz and Krista Lagus. Inducing the morphological lexicon of a natural language from unannotated text. In Proceedings of AKRR, pages 106–113, 2005.

[34] Philipp Koehn. Statistical significance tests for machine translation evaluation. In Proceedings of EMNLP, pages 388–395, 2004.

[35] 麦热哈巴.艾力，姜文斌，王志洋，吐尔根.依布拉音，刘群. 维吾尔语词法分析的有向图模型. 软件学报, 23(12):3115－3129, 2012.

作者简介：
王志洋： 中科院计算所智能信息处理重点实验室、博士研究生 wangzhiyang@ict.ac.cn
吕雅娟： 中科院计算所智能信息处理重点实验室、副研究员
孙 萌： 中科院计算所智能信息处理重点实验室、硕士研究生
姜文斌： 中科院计算所智能信息处理重点实验室、助理研究员
刘 群： 中科院计算所智能信息处理重点实验室、研究员